



Sequence

£710.00 101

PAD
08
3775888P

PROKARYOTIC REVERSE TRANSCRIPTASE

RELATED CASES

MW
9-27-94

1/91 This is a continuation-in-part of prior copending U.S. patent application Serial No. 07/315,427, filed February 24, 1989 and since issued as U.S. Patent No. 5,079,151 on January 7, 1992, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/315,316, filed February 24, 1989 and since issued as U.S. Patent No. 5,320,958 on June 14, 1994, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/315,432, filed on February 24, 1989 and since abandoned, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/517,946, filed on May 2, 1990, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/518,749, filed on March 2, 1990, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/753,110, filed on August 30, 1991, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/817,430, filed January 6, 1992, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/979,447, filed November 20, 1992, respectively which are incorporated herein by reference. *31*

FIELD OF THE INVENTION

The invention relates to bacterial RT enzymes which are capable of synthesizing a hybrid RNA-DNA molecule, called msDNA together with the genes which synthesize the DNA and RNA portion of the molecule.

Another aspect of the invention relates to the isolation and purification of RTs from
20 bacterium which is capable of synthesizing msDNA. The invention deals with groups of prokaryotes

LAW OFFICES
WEISER & ASSOCIATES
SUITE 500
230 SO. FIFTEENTH ST.
PHILADELPHIA, PA 19102
(215) 875-8383
TELECOPIER (215) 875-8394

bacterium capable of synthesizing msDNAs is identified by testing positive by an appropriate screening test.

This is the first time that, as taught in the subject parent patent applications, reverse transcriptase has been found and isolated from a prokaryote.

5

BACKGROUND OF THE INVENTION

Previously, there was described a chromosomal region of the bacterium Myxococcus xanthus which coded for the RNA and DNA portions of an msDNA. Dhundale *et al.* (Dhundale '87) "Structure of msDNA from Myxococcus xanthus: Evidence for a Long, Self-Annealing RNA precursor for the Covalently Linked, Branched RNA", Cell, Vol. 51, pages 1105-1112 (December 24, 1987). Dhundale *et al.* speculated that an Alu I nucleotide fragment contained all the essential coding regions to produce an msDNA. This speculation turned out to be in error.

The Alu I fragment of Dhundale *et al.*, in fact, and inherently did not contain the gene sequence coding for an RT. The Alu I fragment was too short to code for the gene sequence coding for an RT. This was proven by way of sequence analysis by a computer program which searches for open reading frames that can potentially code for a protein. The print-out of the sequence analysis clearly shows that there is no translational reading frame in the Dhundale *et al.* fragment open across a stretch of DNA sufficiently long enough to encode any reverse transcriptase.

What is reported in Dhundale *et al.* in 1987 with respect to a bacterial reverse transcriptase was totally contrary to accepted dogma at that time about the distribution of these enzymes, i.e., that they were present only in viruses which infect eukaryotic organisms.

For the 20 years since the discovery of reverse transcriptase, it was believed that these enzymes were restricted to viruses which infect eukaryotic cells. Now, in accordance with the invention, reverse transcriptases have been identified in bacteria.

SUMMARY OF THE INVENTION

In accordance with the invention, it is shown that various bacteria have nucleotide sequences named "retrons" which encode reverse transcriptases (RTs) which are capable of synthesizing msDNAs. The invention also relates to the isolated and purified bacterial RTs. It has 5 also been determined that the RTs of the bacteria which synthesize msDNAs possess common conserved nucleotide sequences and amino acid residues.

Representative members of the Enterobacteriaceae, Rhizobiaceae and Mycobacteriaceae families are demonstrated to be capable of synthesizing msDNA. These bacteria can be screened for the capability of synthesizing msDNA by an RT labeling or extension in vitro test.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the restriction map of the 3.4 kb fragment around msd and downstream of msr.

Figure 2 shows the nucleotide sequence of the chromosomal region encompassing the msDNA and msd RNA coding regions and an ORF region downstream of msr and the amino acid sequence of Mx162-RT.

Figure 3 shows the amino acid sequence alignment of the msDNA-Mx162 ORF with a portion of the retroviral Pol sequences from HIV and HTLV1 and the ORF of msDNA-Ec67.

Figure 4 shows the sequence similarity of the msDNA-Mx162 reverse transcriptase 20 with other retroelements.

Figure 5 shows the sequence comparison of the regions around the YXDD box of various reverse transcriptases.

Figure 6 shows the detection of msDNA in a clinical isolate of E. coli.

Figure 7 shows the complete primary and proposed secondary structure of msDNA-Ec67.

Figure 8 shows the determination of the RNA nucleotide sequence for the branched RNA linked to msDNA.

5 Figure 9 shows the southern blot analysis of E. coli Cl-1 Chromosomal DNA(A) and analysis of msDNA synthesis by pCl-1E and pCl-1P(B).

Figure 10 shows the restriction map of the 11.6 kb Eco RI fragment.

10 Figure 11 shows the nucleotide sequence of the region from the E. coli Cl-1 chromosome encompassing the msDNA, msd RNA and ORF coding regions and the amino acid sequence of Ec67-RT.

Figure 12 shows the amino acid sequence alignment of the E. coli msDNA ORF with a portion of the retroviral Pol sequence from HIV and HTLV1.

Figure 13 shows the detection of RT activity from various cell extracts.

Figure 14 shows the amino acid sequence alignment of bacterial RTs.

Figure 15 shows the nucleotide and amino acid sequence of Mx65-RT.

Figure 16 shows the nucleotide and amino acid sequence of Sa163-RT.

Figure 17 shows the nucleotide and amino acid sequence of Ec73-RT.

Figure 18 shows the nucleotide and amino acid sequence of Ec86-RT.

Figure 19 shows the nucleotide and amino acid sequence of Ec107-RT.

20 Figure 20 shows the msDNAs from total RNA prepared from each bacterial strain were specifically labeled with ³²P by the RT extension method (12, 14).

Figure 21 shows a collection of 63 rhizobial isolates screened for the presence of msDNA by the RT extension method.

DETAILED DESCRIPTION OF THE DRAWINGS

Figure 1. Restriction Map of the 3.4-kb fragment Around msd and Downstream of msr.

The locations and the orientation of msDNA and msdRNA are indicated by a small 5 arrow and an open arrow, respectively. A large solid arrow represents an ORF and its orientation. The only two AluI sites (A and B) are shown and the DNA sequence between AluI (A) and AluI (B) was determined previously by Yee et al. (1984).

Figure 2. Nucleotide Sequence of the Chromosomal Region Encompassing the ^{See, ID NO. 1 and Ser ID NO: 2} msDNA and msdRNA Coding Regions and an ORF Region Downstream of msr.

The upper strand beginning at the Alu I (A) site (see Figure 1) and ending just beyond the ORF is shown. Only a part of the complementary lower strand is shown from base -301 to -600. The boxed region of the upper strand (332-408) and the boxed region of the lower strand (401-562) correspond to the sequences of msdRNA and msDNA respectively (Dhundale et al., 1987). The starting sites for DNA and RNA and the 5' to 3' orientations are indicated by open arrows. The msdRNA and msDNA regions overlap at their 3' ends by 8 bases. The circled G residue at position 351 represents the branched rG of RNA linked to the 5' end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences proposed to be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA (Dhundale et al., 1987). The ORF begins with the initiation codon at base 640. Single letter designations are 20 given for amino acids. The YXDD amino acid sequence highly conserved among known RT proteins is boxed. Numbers on the right hand column enumerate the nucleotide bases and numbers with a* enumerate amino acids. Small vertical arrows labeled Alu I and SmaI locate the Alu I and SmaI restriction cleavage sites, respectively. The DNA sequence was determined by the chain termination method (Sanger et al., 1977) using synthetic oligonucleotides as primer.

Figure 3. Amino acid Sequence Alignment of the msDNA-Mx162 ORF with a
 Portion of the Retroviral Pol Sequences from HIV and HTLV1 and the ORF of msDNA-Ec67.

Amino acid sequences are compared with matching residues assigned as follows: (o) amino acid residues shared by all four proteins; (o) amino acid residues shared by msDNA-Mx162 and msDNA-Ec67 RTs; (x) amino acid residues shared by msDNA-Mx162 RT with HIV or HTLV1 RTs. Amino acid sequences showed are from residue-177 to -439 for HIV RT (Ratner *et al.*, 1985); residue-15 to -277 for HTLV1 RT (Seiki *et al.*, 1983); residue-32 to -291 for Ec-67 RT (Lampson *et al.*, 1989); and residue-170 to -435 for Mx-162 RT (this work). The YXDD consensus sequence is outlined with a box.

Figure 4. Sequence Similarity of the msDNA-Mx162 Reverse Transcriptase with Other Retroelements. A. Sequence similarity of the region from residue-18 to -128 of the msDNA-Mx162 RT (see Figure 2) with a carboxyl terminal region of integrase of Moloney murine leukemia virus (Mo-MLV) (residue-1070 to -1179; Shinnick *et al.*, 1981). B. Comparison of the sequence from residue-411 to -485 of the msDNA-Mx162 RT (see Figure 2) with the sequence from residue-396 to -461 of the gap protein of human immunodeficiency virus (HIV; Ratner *et al.*, 1985).

Figure 5. Sequence Comparison of the Regions Around the YXDD Box of Various Reverse Transcriptases.

The region from residue-304 to residue-371 of the msDNA-Mx162 RT (see Figure 2) is aligned with various RTs from different sources. The identical amino acid residues with the msDNA-Mx162 RT are indicated by open circles. The YXDD sequences are boxed. The residue numbers for the amino terminal residues and for the carboxyl terminal residues are indicated by the left and the right hand sides of the sequences, respectively. Mx-162 RT from this work (Figure 2); Ec-67 RT from Lampson *et al.* (1989); Ec-86 RT from Lim and Maas (1989); HIV RT from Ratner *et al.* (1985); HTLV1 RT from Seiki *et al.* (1983); Mo-MLV RT from Shinnick *et al.* (1981); RSV (Rous sarcoma virus) RT from Dickson *et al.* (1982); BLV (bovine leukemia virus) RT from Rice

Seq. ID No. 18

et al. (1985); Mt. plasmid (Neurospora mitochondrial plasmid) RT from Nargang et al. (1984); 17.6

Seq. ID No. 19

Drosophila retrotransposon from Saigo et al. (1984); gypsy Drosophila retrotransposon from Yuki et

al. (1986); Tal-3 plant (Arabidopsis thaliana) retrotransposon from Voytas and Ausubel (1988); and

Seq. ID No. 20

Seq. ID No. 23

Ty912 yeast retrotransposon from Clare and Farabaugh (1985). Small arrows in Copia, Tal-3 and

Seq. ID No. 21

Seq. ID No. 22

Ty912 indicate positions of insertions of extra sequences of 18, 18 and 13 residues, respectively. B,

Phylogenetic relationships among various RTs listed in A. The branching positions are arbitrarily

illustrated.

Figure 6. Detection of msDNA in a clinical isolate of E. coli. Total RNA, prepared (Maniatis et al., 1982) from a 5-ml culture, was added to 50 μ l of a reaction mixture containing: 50 mM Tris-HCl (pH8.3); 6 mM MgCl₂; 40 mM KCl; 5 mM DTT; 1 μ M dATP, dTTP, and dGTP; 0.04 μ M dCTP; 0.2 μ M [α -³²P]dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture was incubated at 37°C for 30 min. followed by extraction with 50 μ l phenol-chloroform (1:1) and ethanol precipitation. The samples were electrophoresed on a 4% acrylamide - 8 M urea gel. Lanes: (S) molecular weight markers; MspI digest of pBR322 end-labeled with [α -³²P]dCTP and the Klenow fragment of DNA polymerase I, (1) E. coli K-12 strain C600, (2) the same as in lane 1 except the sample was treated with RNase A (5 μ g, 10 min at 37 °C) just prior to electrophoresis, (3) clinical isolate Cl-1, (4) clinical isolate Cl-1 treated with RNase A. The clinical isolate was identified as Escherichia coli (The clinical E. coli strains were urinary tract isolates kindly provided by Dr. Melvin Weinstein from the microbiology laboratory, R.W. Johnson Hospital, New Brunswick, NJ. The clinical strain Cl-1 was identified using the API-20E identification system (API laboratory products) and gave a typical E. coli profile number of 5044552).

Figure 7. The complete primary and proposed secondary structure of msDNA-

Ec67. The DNA sequence was determined by the Maxam and Gilbert method (Maxam et al., 1980)

using 3'-end labeled msDNA. The RNA sequence (msdRNA; boxed region) was determined using

base-specific RNases as previously described (Dhundale et al., 1987). The 2',5' Branched linkage

between the 15th rG residue and the 5' end of the DNA strand was determined using the debranching enzyme from HeLa cells as described previously (Dhundale *et al.*, 1987; Furuichi *et al.*, 1987; Ruskin *et al.*, 1985; Arenas *et al.*, 1987; the debranching enzyme was a gift from Jerard Hurwitz). The branched rG at position 15 is circled, and both RNA and DNA are numbered from their 5' ends.

5 Figure 8. Determination of the RNA nucleotide sequence for the branched RNA
linked to msDNA. Total RNA was prepared from the clinical strain Cl-1 and fractionated on a 5%
acrylamide gel. msDNA containing full length RNA was eluted from the gel. This fraction was then
labeled at the 5' end of the RNA with [γ -³²P]ATP and T4 polynucleotide kinase. The 5' end labeled
(piece of SEQ ID NO:25)
RNA linked to msDNA was again purified on an 18% acrylamide - 8M urea sequencing gel. The
labeled RNA was then sequenced using limited digestion with base-specific RNases as described
previously (Dhundale *et al.*, 1987). Lanes: OH⁻, partial alkaline hydrolysis ladder; (0.5 M sodium
bicarbonate/carbonate pH9.2); -E, no enzyme treatment of the labeled RNA linked to msDNA; T1,
RNase T1 (1U/reaction, 55°, 15 min.); U2, RNase U2 (1U and 0.5U/reaction, 55°, 15 min.); PhyM,
RNase PhyM (1U/reaction, 55°, 15 min); Bc, RNase B. cerus (2U/reaction, 55°, 15 min.); CL3, RNase
CL3 (2U/reaction, 37°, 15 min.). The large gap in the sequence gel is due to msDNA linked at the
rG residue at position 15 by a 2',5' phosphodiester linkage (Furuichi *et al.*, 1987). The RNA sequence
at the 3'-end region from the branched rG residue (the upper part of the gel) was determined from
6% gel (data not shown).

Figure 9. Southern blot analysis of *E. coli* Cl-1 chromosomal DNA(A) and analysis of
msDNA synthesis by pLI-1E and pCl-1P(B). A: The chromosomal DNA was digested with EcoRI
(lane 1), HindIII (lane 2), BamHI (lane 3), PstI (lane 4), and BglIII (lane 5). For each lane, 3 μ g of the
DNA digest was applied to a 0.7% agarose gel. After electrophoresis the gel was blotted to a
nitrocellulose filter, and hybridization analysis was carried out according to Southern (Southern, 1975)
using msDNA labeled by AMV-RT with $[\alpha-32\text{P}]$ dCTP as a probe. Numbers at the left represent the
molecular weights in kb. B: Total DNA prepared from each strain was treated with RNase A,

LAW OFFICES
WEISER & ASSOCIATES
SUITE 500
230 SO. FIFTEENTH ST.
PHILADELPHIA, PA 19102
(215) 875-8383
TELECOPIER (215) 875-8394

separated on a 5% acrylamide gel and stained with ethidium bromide. Lane S, pBR322 digested with MspI used for molecular size markers; lane 1, DNA prepared from the host strain CL-83(recA⁻); lane 2, CL-83 (recA⁻) transformed with plasmid pCl-1E (11.6 kb EcoRI fragment; see Figure 5); lane 3, with plasmid pCl-1P (2.8- kb PstI(a)-PstI(b) fragment; see Figure 5). An arrow indicates the position 5 of msDNA.

Figure 10. Restriction map of the 11.6-kb EcoRI fragment. In the Cl-1E map, the left-hand half (EcoRI to HindIII) was not mapped. In the Cl-1EP5 map, the locations and the orientations of msDNA and msdRNA are indicated by a small arrow and an open arrow, respectively. A large solid arrow represents an ORF and its orientation.

Figure 11. Nucleotide sequence of the region from the E. coli Cl-1 chromosome encompassing the msDNA and msdRNA coding regions and an ORF downstream of the msdRNA region. The entire upper strand beginning at the BalI site (see Figure 5) and ending just beyond the ORF is shown. Only a part of the complementary lower strand is shown from base 241 to 420. The long boxed region of the upper strand (249-306) corresponds to the sequence of the branched RNA (msdRNA; see Figure 7) portion of the msDNA molecule. The boxed region of the lower strand corresponds to the sequence of the DNA portion of msDNA (see Figure 7). The starting site for DNA and RNA and the 5' to 3' orientations are indicated by large open arrows. The msdRNA and msDNA regions overlap at their 3' ends by 7 bases. The circled G residue at position 263 represents the branched rG of RNA linked to the 5' end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences proposed to be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA (Dhundale *et al.*, 1987). Note that the nucleotide at position 257 (U on the RNA transcript) and the nucleotide at position 373 (G on the RNA transcript) form a U-G pair in the stem between sequence a1 and a2. The proposed promoter elements (-10 and -35 regions) for the primary RNA transcript are also boxed. The ORF begins with the initiation codon at base 418. Single letter designations are given for amino acids. The YXDD

amino acid sequence conserved among known RT proteins is boxed. Numbers on the right hand column enumerate the nucleotide bases and numbers with a* enumerate amino acids. Small vertical arrows labeled H and P locate the HindIII and PstI restriction cleavage sites, respectively. The DNA sequence was determined by the chain termination method (Sanger *et al.*, 1977) using synthetic oligonucleotides as primers.

5 *Seq. FD No 29*
Seq. FD No 29 *Seq. FD No 29*
G B Figure 12. Amino acid sequence alignment of the *E. coli* msDNA ORF with a portion of the retroviral Pol sequence from HIV and HTLV1. Amino acid sequences are compared with matching residues assigned as follows: (+) amino acid common to msDNA and HIV RTs; (o) amino acid shared by msDNA and HTLV1 RTs; and (o) amino acid shared by all three proteins. Arrows divide the protein sequences into three functional domains (Toh *et al.*, 1983; Geng *et al.*, 1985; Varmus, 1985, Tanese *et al.*, 1988): An amino terminal RT domain, a carboxy terminal RNase H region, and a central "tether" region. The specific amino acid residues for the RT, tether, and RNase H domains, for each protein are: HIV, 177-439, 440-600, 601-722 respectively; HTLV1, 15-277, 278-462, 463-592 respectively; msDNA ORF, 32-290, 291-465, 466-586 respectively. The YXDD polymerase consensus sequence is outlined with a box.

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185

DEAE paper (DE81; Whatman Inc.). The paper was washed to remove unincorporated label (Tanese *et al.*, 1985; Tanese *et al.*, 1986) and then exposed to an X-ray film. In row (A) all reactions contain added template primer (poly rC-dG). Row (B) contains control reactions in which no template-primer is added. Columns contain the designated cell extracts: M-MuLV, cloned Moloney Murine Leukemia Virus RT gene; pGB2 (Churchward *et al.*, 1984), vector plasmid in strain C2110; pCl-1EP5, recombinant plasmid with the cloned msDNA gene. The large amount of background activity observed with the M-MuLV control extract is due to the activity of DNA Polymerase I since this extract is obtained from a PolA⁺ strain (HB101).

Figure 14 shows the amino acid sequence alignment of bacterial RT carried out according to Xiong and Eickbush (1990). Amino acids highly conserved in eukaryotic RTs are shown at the top of the sequences. These amino acids include largely unvaried residues or chemically similar residues. (h) Hydrophobic residue; (p) small polar residues; (c) charged residue. Amino acids conserved in all seven bacterial RTs (identical residues plus functional conserved residues indicated by h for hydrophobic residues or p for polar residues) are marked by solid dots at the bottom of the sequences. The consensus sequence shown at the bottom of the sequences is determined when five out of seven sequences contain an identical or a chemically similar residue (h, hydrophobic residue; p, charged and polar residue). The subdomains 1 to 7 are according to Xiong and Eickbush (1990), which are boxed and indicated by numbers. The highly conserved YXDD sequences are also boxed.

Numbers on the right indicate the amino acid positions from the amino terminus for each RT.

Sources for the sequences are Sal63 (Hsu *et al.* 1992b), Mx162 (Inouye *et al.* 1989), Mx65 (Inouye *et al.* 1990), Ec67 (Lampson *et al.* 1989b), Ec86 (Lim and Maas 1989), Ec73 (Sun *et al.* 1991), and Ec107 (Herzer *et al.* 1992).

Figure 15 shows nucleotide sequence of the chromosomal region encompassing the Mx65-msDNA and msdRNA coding regions and an ORF region downstream of msr. The sequence covers from the Alu I(A) site to 78 bp downstream of the ORF. The complementary strand is only

shown from bases 121-300. The boxed region of the upper strand (positions 143-191) and the boxed region of the lower strand (positions 186-250) correspond to the sequences of msdRNA and msDNA, respectively. The starting sites for DNA and RNA and the 5' to 3' orientation are indicated by open arrows. The msdRNA and msDNA regions overlap at their 3' ends by 6 bases. The circled G residue at position 206 represents the branched guanosine of RNA linked to the 5' end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences proposed to be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA. The ORF begins with the initiation codon at base 279. The YXDD amino acid sequence highly conserved among known RT proteins is boxed. Numbers on the right-hand column enumerate the nucleotide bases, and numbers with asterisks enumerate amino acids (single-letter code). The DNA sequence was determined by the chain-termination method using synthetic oligonucleotides as primers.

Figure 16 shows nucleotide sequences of 3,060 bases encompassing msr, msd, and the RT gene of S. aurantiaca. The sequence from base 421 to base 720 which contains msr and msd is shown double stranded. The boxed regions of the upper strand (bases 440 to 540) and the lower strand (bases 508 to 670) correspond to the sequences of msdRNA and msDNA, respectively. The starting sites for msDNA and msdRNA are indicated by open arrows. The circled G at the position 458 is the branched rG of msdRNA linked to the 5' end of msDNA. Long solid arrows labeled with a1 and a2 represent inverted repeated sequences proposed to form the secondary structure in the primary RNA transcript which serves to prime msDNA synthesis. Amino acids are indicated by single letters. The YXDD sequence highly conserved among known RTs is boxed. X^e and B^f sites are indicated by arrows. Numbers on the right-hand side and numbers with asterisks represent numbers for bases and amino acids, respectively.

8ED ID NO: 43 8ED ID NO: 44 and 8ED ID NO: 45
Seq. ID No. 39

Figure 17 shows the sequences of msdRNA and msDNA which are boxed and their orientations are indicated by open arrows. The branched G residue at position 10425 is circled. The

inverted repeat sequences require for the biosynthesis of msDNA - Ec73 are shown by arrows labeled a1 and a2. Amino acid residues of Ec73-RT are shown by a single-letter code put at the center of each codon. *Seq. ID No. 40*

Figure 18 shows the restriction map of the 3.5 kb insert of pDB808 and nucleotide sequence of chromosomal determinants of the msDNA-RNA compound of E. coli B. (A) Restriction map of the 3.5 kb insert of clone pDB808. The solid bar represents the region whose sequence is presented in (B). Transcription is from left to right. Restriction enzymes are: P, PstI, H, HpaI; B, BgIII; X, Xhol. (B) Nucleotide sequences of the chromosomal determinants. Only the strand corresponding to the transcript is shown. Nucleotides are numbered starting from the first base observed in the msdRNA. The msdRNA coding region is overlined, and the msDNA coding region is underlined. The msDNA sequence is complementary to the sequence shown in this figure. Inverted repeats are indicated by double-dashed lines. The G at position 14 is the branched guanylate of msdRNA in the msDNA-RNA compound. IR, 12 bp inverted repeat. *Seq. ID No. 44, and Seq. ID No. 47*

Figure 19 shows sequence of the retron and flanking regions of Ec107. The sequences corresponding to the K-12 genomic DNA are shown in lower case letters from bases 1-99 and 1400-1540. The msRNA and msDNA regions are boxed. Also indicated are the a1-a2 conserved inverted repeats (indicated by arrows) and the branched G, which is circled. The RT consists of 319 amino acids and contains the YXDD sequence (boxed) which is highly conserved among known RTs. The transcription start site occurs at base 170; a possible terminator is indicated by head-to-head arrows following the RT coding region. Primer extension was utilized in order to determine the transcription start site. These sequence data will appear in the EMBL/GenBank/DDJB Nucleotide Sequence Data Libraries under the accession number X62583. *48 and Seq II*
Seq. ID No. 42, NC:49

DETAILED DESCRIPTION OF THE INVENTION

The description which follows describes msDNA and RT from Myxococcus xanthus.

This is a typical bacterium which belongs to a genus of bacteria, whose representative members possess an RT capable of synthesizing msDNA.

5 The existence of a peculiar branched RNA-linked DNA molecule called msDNA (multicopy single-stranded) has been demonstrated in various myxobacteria, Gram-negative soil bacteria (Yee et al., 1984; Dhundale et al., 1985; Furuichi et al., 1987a,b; Dhundale et al., 1987; Dhundale et al., 1988b). msDNA (msDNA-Mx162) from Myxococcus xanthus consists of 162-base single stranded DNA, the 5' end of which is linked to the 2' position of the 20th rG residue of a 77-base RNA molecule (msdRNA) by a 2', 5'-phosphodiester linkage (Dhundale et al., 1987). It exists at a level of approximately 700 copies per genome. Stigmatella aurantiaca also possesses an msDNA (msDNA-Sal63) which is highly homologous to msDNA-Mx162 (Furuichi et al., 1987b). In addition to msDNA-Mx162, M. xanthus has another smaller species of msDNA (mrDNA or msDNA-Mx65), which has no primary sequence homology with msDNA-Mx162 or msDNA-Sal63 (Dhundale et al., 1988b). However, all msDNAs so far characterized share key structural features such as a branched rG residue, stem-and-loop structures in RNA and DNA molecules, and a DNA-RNA hybrid at the 3' ends of DNA and RNA molecules.

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

20

Previously it was predicted that reverse transcriptase is required for msDNA biosynthesis on the basis of the finding that msdRNA is derived from a much longer precursor, which can form a very stable stem-and-loop structure (Dhundale et al., 1987). This precursor molecule was proposed to serve as a primer for initiating msDNA synthesis as well as a template to form the branched RNA-linked msDNA. The latter reaction requires reverse transcriptase activity. In M. xanthus, the region coding for the RNA molecule (msr) is located on the chromosome in the opposite orientation to the msDNA coding region (msd) with the 3' ends overlapping by 6 bases for msDNA-Mx65 (Dhundale et al., 1988b) or by 8 bases for msDNA-Mx162 (Dhundale et al., 1987). In addition, as in all the msDNAs found in myxobacteria, there is an inverted repeat comprised of a 14-base

sequence for msDNA-Mx65 (Dhundale *et al.*, 1988b) or a 34-base sequence for msDNA-Mx162 (Dhundale *et al.*, 1987) and a 33-base sequence for msDNA-Sal63 (Furuichi *et al.*, 1987b) immediately upstream of the branched G residue and a sequence immediately upstream of the msDNA coding region. As a result of this inverted repeat, a longer primary transcript beginning upstream of the RNA coding region and extending through the msDNA coding region is considered to self-anneal and form a stable secondary structure. When three base mismatches were introduced into the secondary structure immediately upstream of the branched rG residue, msDNA synthesis was almost completely blocked. However, if three additional base substitutions were made on the other strand to resume the complementary base pairing, msDNA production was restored (Hsu *et al.*, 1989). This result strongly supports the proposed model for msDNA synthesis.

It was also shown that a deletion mutation at the region 100 base pairs (bp) upstream of the DNA coding region (msd) and an insertion mutation at a site 500 bp upstream of msd caused a significant reduction in msDNA production (Dhundale *et al.*, 1988a). This indicates that there is a cis- or trans-acting positive element required for msDNA synthesis in this region. In this report we determined the DNA sequence of this region and found an opening reading frame (ORF) coding for 485 amino acid residues beginning with an initiation codon, ATG, which is located 77 bp upstream of msd (or 231 bp downstream of msr). The very close proximity between msd and the ORF suggests that they may be transcribed as a single transcript. The amino acid sequence of the ORF shows similarity with retroviral reverse transcriptases. We discuss a possible origin of the reverse transcriptase gene as well as a possible relationship between the msDNA system and retroviruses. Recently, some strains of Escherichia coli were found to produce msDNA and the gene for reverse transcriptase which is essential for msDNA production, is linked to the msd region, (Lim and Maas, 1989; Lampson *et al.*, 1989b). Comparison of the msDNA systems of M. xanthus and E. coli raises an intriguing question as to how the extensive diversity found in msDNA systems has emerged in bacteria and what possible functions msDNA may have.

In a preceding paper, it was demonstrated that msDNA is in fact synthesized by reverse transcriptase in a cell-free system in M. xanthus (Lampson *et al.*, 1989a).

Reverse transcriptases are isolated, and if desired, purified, and biological characterization carried out, if desired, by known methods such as those described in Lampson, B.C., M. Viswanathan, M. Inouye and S. Inouye, "Reverse Transcriptase from Escherichia coli Exists as a Complex with msDNA and is Able to Synthesize Double-stranded DNA", J. Biol. Chem. 265: 8490-8496 (1990), which is incorporated by reference as if fully set forth herein.

RESULTS AND DISCUSSION

Identification of an ORF Associated with msd

On the basis of mutations closely associated with msd which significantly reduce msDNA production, it was assumed that in this region there is a cis- or trans-acting element which is essential for msDNA synthesis (Dhundale *et al.*, 1988a). Figure 1 shows a restriction map around msd. The msDNA coding region is shown by a thin arrow from right to left (msd), and the msdRNA coding region by a thick open arrow (msr). In the previous work (Dhundale *et al.*, 1988a), two mutations were constructed; one, a deletion mutation in which the sequence from Alu I(b) to SmaI was replaced by a gene for kanamycin resistance (see Figure 1), and the other an insertion mutation at the SmaI site by a gene for kanamycin resistance (see Figure 1).

20

In order to elucidate the properties of the element required for msDNA production, the DNA sequence of the region upstream of msd was determined as shown in Figure 2. A long open reading frame (ORF) beginning with an initiation codon was found 77 bases upstream of msd. The ORF is preceded by a ribosome binding sequence of AGG (residue 630 to 632) 7 bases upstream of the initiation codon. The ORF codes for a polypeptide of 485 amino acid residues. The Alu I(b) and SmaI sites (see Figure 1), where mutations inhibiting msDNA synthesis were created, are located at amino acid residue -12 and -142 of the ORF, respectively or at the nucleotide sequence from residue -672 to -675, and from residue -1061 to -1066, respectively (Figure 2). In Figure 2, msd or the DNA sequence corresponding to the msDNA sequence is indicated by the closed box on the lower strand

and the orientation is from right to left. Similarly, the msdRNA sequence (msr) is also indicated by

the closed box on the upper strand and the orientation is from left to right. The msd and msr regions overlap by 8 bases. An inverted repeat is also indicated by arrows with letters a1 and a2. This inverted repeat comprises a 34-base sequence immediately upstream of the branched G residue (residue 317 to 350; sequence a2 in Figure 2) and another 34-base sequence at the 3' end (residue 597 to 564; sequence a1). This inverted repeat is essential to form a stem structure which provides a stable secondary structure in a long primary transcript. This secondary structure is considered to serve as the primer as well as the template for msDNA synthesis (Dhundale *et al.*, 1987; Hsu *et al.*, 1989).

Sequence Similarity with Retroviral Reverse Transcriptases

When the amino acid sequence of the ORF was compared with known proteins, a striking similarity was found between the sequence from Leu-308 to Ser-351 and retroviral reverse transcriptases (RT). In particular, this region contains the YXDD sequence, the highly conserved sequence in all known RTs. This sequence (Tyr-344 to Asp-347) is boxed in Figure 2. In Figure 3, the ORF sequence of 266 amino acid residues from Ala-170 to Lys-435 is compared with RTs from HIV (human immunodeficiency virus; Ratner *et al.*, 1986) and HTLV1 (human T-cell leukemia virus type 1; Seiki *et al.*, 1983). As mentioned above, within the sequence of 44 amino residues from Leu-308 to Ser-351, there are 14 and 12 identical residues with HIV (32%) and HTLV1 (27%), respectively. The entire RT domains of HIV and HTLV can also be aligned with the ORF sequence from Ala-170 to Lys-435, with much less similarity as shown in Figure 3. However, the same region was found to be extremely well aligned with the RT which was recently found in a clinical strain of Escherichia coli (Lampson *et al.*, 1989b). This E. coli RT consists of 586 amino acid residues, and its amino terminal domain (residue-32 to -291) and the carboxyl terminal domain (residue-466 and -586) have been demonstrated to have sequence similarity with retroviral RT and ribonuclease H. This RT gene from E. coli was shown to be required for the production of msDNA (msDNA-Ec67) and to have reverse transcriptase activity (Lampson *et al.*, 1989b). Figure 3 shows that the sequence similarity between E. coli and M. xanthus RTs is distributed within almost the entire RT region; in particular in the region from Tyr-181 to Ser-212, 15 out of 32 residues are identical (47% similarity);

in the region from Gly-226 to Gly-265, 19 out of 40 residues (48% similarity); in the region from Leu-308 to Ser-351, 26 out of 44 residues (59% similarity); and in the region from Lys-354 to Asn-408, 21 out of 55 residues (38% similarity). Overall, similarity from Ala-170 to Lys-435 is 32% (85 out of 266 residues are identical). In spite of these similarities, the M. xanthus ORF does not have 5 the domain, which shows apparent sequence similarity with ribonuclease H (RNase H). The RNase H domain is found to be located in the carboxyl terminal region of the same polypeptide in which the RT domain exists in the amino terminal region in the case of the E. coli RT and other retroviral RTs. In the preceding paper, it was shown that there is a precise coupling between RT and RNase H activity (Lampson *et al.*, 1989a). Therefore, RNase H may still reside with the ORF, or RNase H may 10 be encoded by a separate gene.

Sequence Similarity with Other Proteins

In contrast to the E. coli RT and other retroviral RTs, the ORF found in M. xanthus has a long amino terminal extra domain consisting of approximately 170 residues. Interestingly, this region shows some sequence similarities with the carboxyl terminal region associated with integration protein of Mo-MLV (Moloney murine leukemia virus; Shinnick *et al.*, 1981) (see Figure 4A); the sequence from Pro-18 to Leu-128 of the ORF shows 22% similarity (24 out of 111 residues) with the region from Pro-1070 to Leu-1179 of the gag-pol polyprotein of Mo-MLV. It should be noted that this region of Mo-MLV is unique for Mo-MLV integration protein and does not share sequence similarity with other retroviral endonucleases (Johnson *et al.*, 1986). It is also interesting to notice 20 that in Ty retrotransposon, this domain is located in front of the RT domain in contrast to the retroviral endonuclease domain (Clare and Farabaugh, 1985).

As pointed out above, the ORF does not have homology to E. coli or retroviral RNase H. Instead, it has a short sequence of approximately 80 residues after the RT domain. In this region, one can also find sequence similarity with a part of the gag region of HIV. As shown in Figure 4B, the sequence from Gly-411 to Glu-485 has 22 identical amino acid residues (31% similarity) with the 25 region from Gly-396 to Pro-461 of the gag protein of HIV (Ratner *et al.*, 1985).

Requirement of Reverse Transcriptase

The fact that disruption of the ORF significantly reduced msDNA production in M. xanthus (Dhundale *et al.*, 1988a) and the fact that the ORF has sequence similarity with retroviral RTs strongly supports the previous hypothesis that RT is required for the synthesis of msDNA (Dhundale *et al.*, 1987). Recently, we were able to demonstrate that msDNA is indeed synthesized by reverse transcriptase activity in a cell-free system (Lampson *et al.*, 1989a). The fact that a small amount of msDNA (3% of the wild type level) is still produced in the ORF mutants (Dhundale *et al.*, 1988a) is most likely due to another RT associated with smaller msDNA (msDNA-Mx65; previously assigned mrDNA; Dhundale *et al.*, 1988b). In fact, an ORF has been found to be associated with the region responsible for msDNA-Mx65 production.

At present it is unknown if the ORF is transcribed together with msdRNA from a common upstream promoter or if the ORF has its own independent promoter. Previously, a major RNA transcript of approximately 375 bases by S1 mapping (Dhundale *et al.*, 1987) was identified. This transcript covers the region from approximately 75 bases upstream of msr (at around residue-256 in Figure 2) to approximately 70 bases upstream of msd (at around residue-632 in Figure 2). This indicates that this RNA transcript ends at the ribosome binding site (AGG, 630-632) of the ORF. It is possible that the primary RNA transcript covers not only the msr-msd region but also the entire ORF. This transcript of approximately at least 2 kilobases (kb) is then used as the mRNA for the ORF to produce RT. At the same time, the 5' untranslated region of 350 bases forms a stable secondary structure which serves as a primer and a template for msDNA synthesis as previously proposed (Dhundale *et al.*, 1987). Because of the secondary structure, the 5' end region is probably much more stable than the ORF mRNA region. As a result, only the 375-base RNA from the 5' end of the transcript was detected in the previous work. In E. coli, the RT gene was shown to be transcribed from a single promoter for the msr region (Lampson *et al.*, 1989b).

Evolution of Reverse Transcriptase

All of the RTs so far identified are from eukaryotic origins, and associated with either retroviruses or retrotransposons. DNA synthesis for retroviruses and transposition events for retrotransposons occur via RNA which is used as a template for RTs (see review by Varmus, 1985).
5 From amino acid similarity in various RTs, possible evolutionary relationships among these RTs has been proposed (Yuki *et al.*, 1986).

SEARCHED INDEXED
H5
SERIALIZED FILED

The present invention demonstrates that RTs are not specific to eukaryotes but exist in prokaryotes as well. An intriguing question arises as to the evolutionary relationship between prokaryotic and eukaryotic RTs and the origin of RT. In order to compare the amino acid sequences 10 of these RTs, the sequence of the M. xanthus RT from Gly-304 to Leu-371 was chosen, since this sequence includes the YXDD box, the most conserved region among different RTs. In Figure 5A this sequence is compared with 13 other representative RTs from bacteria, yeast, plant, mitochondrial plasmid, and animal retroviruses. Within these 14 sequences, the D-D sequence (residues-346 and -347) is completely conserved, and both G-311 and Y-344 are also well conserved except for Ty-RT. Besides these residues, L-308, P-309, Q-310, S-315, P-316, L-330, S-351, and L-371 are fairly well conserved among these sequences. On the basis of the numbers of identical amino acid residues, M. xanthus RT has the following similarities with other RTs: 47% (32 amino acid residues) with E. coli C1-1 RT; 41% (28) with E. coli B RT; 24% (16) with HIV, BLV, and mitochondrial plasmid RTs; 22% (15) with Mo-MLV RT; 21% (14) with RSV, 17.6, gypsy, and Tal-3 RTs; 19% (13) with HTLV1 RT; 20 15% (10) with Ty912 RT; and 9% (6) with Copia RT. On the basis of the phylogenetic relationships among RTs proposed by Yuki *et al.* (1986), and the present data, a dendrogram of homology of various RTs may be constructed as shown in Figure 5B. As proposed earlier (Yuki *et al.*, 1986), modern RTs are composed to two major groups I and II. One group (group II) consists of retrotransposons found in yeast (Ty912), plant (Tal-3), and Drosophila (Copia). Bacterial RTs seem 25 to belong to the other group (group I) together with other retrotransposons from Drosophila such as 17.6 and gypsy, mitochondrial plasmid RT, and retroviral RTs. This indicates that both prokaryotic and eukaryotic RT genes were possibly derived from a single ancestral RT gene.

Origin of the *M. xanthus* Reverse Transcriptase

In addition to the sequence similarity between the *M. xanthus* RT and RTs from retroviruses and retrotransposons, msDNA shares other interesting similarities with retroviruses and retrotransposons; msDNA (synthesis of single-stranded DNA) starts at a site 77 bases upstream of the RT gene and the orientation of DNA synthesis is opposite to the direction of translation of the RT gene. In the case of retroviruses and retrotransposons, single-stranded DNA synthesis proceeds at the 5'-end untranslated region of an RNA molecule which serves as the mRNA for RT as well (Weiss *et al.*, 1985). The orientation of DNA synthesis is also opposite to the direction of translation of the RT gene. In the case of msDNA synthesis an RNA transcript itself serving as a template also serves as a primer by self-annealing to form a stable secondary structure (Dhundale *et al.*, 1987), whereas in the case of retroviruses and retrotransposons tRNAs are recruited from the cell for the priming reaction. At present it is unknown if branched RNA-linked msDNA is the final product of an unknown function or if it is a stable intermediate leading to other products.

Furthermore, it is of great interest whether the *M. xanthus* RT is associated with a complex such as virus-like particles such as those found for yeast Ty1 element (Eichinger and Boeke, 1988). In a preliminary experiment, msDNA of *M. xanthus* exists as a complex with proteins in the cell which sediments as a 22S particle. Characterization of this complex may shed light on questions concerning the relationship between msDNA and retrocomponents as well as the functions of msDNA.

At present, there is no information to support the possibility that msDNA may be a transposable element or an element associated with a provirus (or prophages). It is important to point out that the RT gene from *M. xanthus* appears to be as old as other genomic genes for the following reasons: (a) Nine independent natural isolates of *M. xanthus* from various sites (including Fiji Island and eight different sites in the United States) contained mutually hybridizable msDNA (Dhundale *et al.*, 1985). Since under the same hybridization condition, msDNA-Mx162 did not hybridize with msDNA-Sa163 [which has extensive homology in both DNA and RNA sequences with msDNA-Mx162; Dhundale *et al.*, (1987)], the nine independent strains *M. xanthus* are assumed to contain almost identical msDNA. (b) The codon usage of the Mx-162 RT is almost identical to those found

in other M. xanthus genes (Table 1). M. xanthus is known to have a very high G+C content (70%; Johnson and Ordal, 1968) and as a result, all the genes so far characterized have very high G+C contents at the third positions of codons used; 85.4% for vegA (Komano et al., 1987), 85.7% of ops (Inouye et al., 1983), 87.2% for tps (Inouye et al., 1983), 88.4% for mbhA (Romeo et al., 1986), and 5 93.9% for sigma factor. The average G+C content of the third positions is calculated to be 90.0% for these genes (Table 1). Surprisingly, the G+C content of the third positions of the RT codons is highest among these genes (95.5%; Table 1).

In contrast, the E. coli msDNA system including the RT gene is considered to have been acquired much later in the evolution of E. coli. Reasons for this conclusion include: (a) Only 10 four strains out of 89 independent clinical E. coli strains were found to produce msDNAs (Lampson et al., 1989b). (b) The codon usage of the E. coli RT is significantly different from the general codon usage of E. coli genes obtained from 199 E. coli genes (Maruyama et al., 1986). In particular, out of 62 arginine codons used in the E. coli RT, 40 (65%) use AGA or AGG in contrast to 2.7% for the AGA+AGG usage among all arginine codons in 199 E. coli genes (see Table 1). The AGA and AGG codons are the least used codons in E. coli (Maruyama et al., 1986). In addition to AGA and AGG codons, many other codons, GCC and GCG for Ala, CGU and CGC for Arg, CAG for Gln, GGC and GGA for Gly, CAC for His, AUC and AUA for Ile, UUA, CUU and CUG for Leu, UUC for Phe, CCU and CCG for Pro, UCG for Ser, ACC and ACA for Thr, and GUC for Val. (c) Although the 15 E. coli msDNAs share little sequence homology, they all share the key secondary structures of a branched rG residue, a DNA-RNA hybrid at the 3' ends of the msDNA and msdRNA, and stem-and-loop structures in RNA and DNA strands (Lampson et al., 1989b; Lim and Maas, 1989).

These results clearly demonstrate distinct differences between the msDNA systems of E. coli and M. xanthus. Myxobacteria are common organisms in soil and are found all over the world regardless of climate, and considered to diverge from their nearest bacterial relatives about 2×10^9 20 years ago when the atmosphere became aerobic (see a review by Kaiser, 1986). Since it is reasonable to assume that the M. xanthus RT gene is as old as other genomic genes, the RT gene existed much before eukaryotic cells appeared ($1.5-0.9 \times 10^9$ years ago). The relatedness between various 25

prokaryotic and eukaryotic RTs as shown in Figures 5A and B strongly supports the existence of a single ancestral gene for all RTs. It is possible that such an ancestral RT gene was independently recruited into different systems such as the msDNA system, the retrotransposon system, and the retroviral system. Alternatively, the msDNA system may be a primitive ancestral system from which retrotransposons and retroviruses originated. In this regard, it is intriguing to point out other sequence similarities between the M. xanthus RT-ORF and other retroelements (see Figure 4) other than RT itself as well as the similar mode of initiation of DNA synthesis by RT as discussed earlier.

At present, it is beyond our speculation why the E. coli msDNA systems are so diverged in contrast to the M. xanthus msDNA system and how they were acquired into the genomes of some E. coli strains. However, it should be noted that the E. coli RTs are most related to the M. xanthus RT indicating that they were not derived from eukaryotic origins. Possible origins of retroviruses have been discussed (Temin, 1980). The recent finding of an imposon in a genetic component for a mouse gene also raises an interesting question concerning the evolution of retroelements (Stavenhagen and Robins, 1988). Further characterization of the prokaryotic RTs and the msDNA systems will provide clues to the origins of RT and other retroelements.

EXPERIMENTAL PROCEDURE

DNA Manipulation and Plasmids

DNA manipulation was performed as described by Maniatis *et al.* (1982). The plasmid isolation was as originally described by Birnboim and Dolly (1979). Plasmid pmsSB7 containing the 20 5 kb SalI-BamHI fragment shown between the SalI and BamHI sites of pUC9 (Vieira and Messing, 1982) was used. After the 2.2 kb SalI-SmaI fragment from pmsSB7 was subcloned between the SalI and SmaI sites of pUC9, all RsaI fragments were gel-purified and cloned into pUC9 for DNA sequence.

DNA sequence

DNA sequence was determined by the chain termination method (Sanger *et al.*, 1977) using single-stranded or double-stranded DNA as templates with synthetic oligonucleotides.

Other Material and Methods

5 Restriction enzymes were purchased from either Bethesda Research Laboratories or New England BioLabs. [α -³⁵S] dATP was from Amersham. Sequenase, Version 2.0 Kit was purchased from United States Biochemical Corporation for DNA sequences.

Cyborg program from International Biotechnologies Inc. was used to search sequence homology in GenBank Release 55.

10 Screening of bacteria for retron synthesized msDNAs was performed by the methods of Lampson *et al.* *J. Bacteriol.*, 173:5363-5370 (1991), or Yee *et al.*, *Cell*, 38, 203-209 (1984).

RTs were identified and isolated by the method of Lampson *et al.*, *J. Biol. Chem.*, 265:8490-8496.

msDNA in *Escherichia coli*

The recent serendipitous finding of msDNA (msDNA-Ec86) in *E. coli* B by Dongbin Lim and Werner Maas (D. Lim *et al.*, 1989) prompted a to search for msDNA in other *E. coli* strains. Previously established by Yee *et al.* (T. Yee *et al.*, 1984), msDNA is not found in the common laboratory strain K12, however, to our surprise, it was in a clinical *E. coli* strain isolated from a patient with a urinary tract infection. Fifty independent *E. coli* urinary tract isolates were examined 20 for the presence of msDNA (The clinical *E. coli* strains were urinary tract isolates kindly provided by Dr. Melvin Weinstein from the microbiology laboratory, R.W. Johnson Hospital, New Brunswick, NJ. The clinical strain Cl-1 was identified using the API-20E identification system (API laboratory products) and gave a typical *E. coli* profile number of 5044552.). The screening method involved treatment of total RNA prepared from each strain with (AMV) RT in the presence of [α -³²P]dCTP plus dATP, dTTP, and dGTP followed by polyacrylamide gel electrophoresis. Since msDNA contains

a DNA-RNA duplex structure, the 3' end of the DNA molecule serves as an intramolecular primer and the RNA molecule as a template for RT. When RNA prepared from one of the clinical strains, E. coli Cl-1, was labeled in this manner, two distinct, low molecular weight bands of about 160 bases became labeled with ^{32}P and are shown in Figure 6. If the labeled sample is digested with 5 ribonuclease (RNase) A prior to loading on the gel, a single band corresponding to 105 bases of single-stranded DNA is detected (lane 4). This indicates that both bands in lane 3 contain a single-stranded DNA of identical size. The two labeled bands observed prior to RNase treatment (lane 3) are due to two species of msDNA comprised of a single species of single-stranded DNA linked to RNA molecules of two different sizes. RNA molecules of two different sizes have been observed at 10 the 5' ends of msDNA from myxobacteria in which a precursor molecule contains a longer RNA which is processed into a smaller mature form (Dhundale *et al.*, 1987; Furuichi *et al.*, 1987). Among the 89 clinical isolates screened, three other strains produced msDNA-like molecules of varying size and quantity, suggesting extensive diversity among these molecules. As previously reported (Dhundale, 1985), msDNA was not observed in the E. coli K-12 strain, C600 (lanes 1 and 2, Figure 6).

Nucleotide sequence of msDNA Ec-67

To determine the base sequence of the DNA molecule, the RNA-DNA complex isolated from the clinical strain was labeled at the 3' end of the DNA molecule with AMV-RT and $[\alpha-^{32}\text{P}]$ dATP. By adding dideoxy-CTP, ddTTP, and ddGTP to the reaction mixture, a single labeled 20 adenine is added to the 3' end of the DNA molecule. RNA is removed with RNase A+ T1 and the end-labeled DNA is subjected to the Maxam and Gilbert sequencing method (Maxam *et al.*, 1980). Figure 7 shows that msDNA consists of a single-stranded DNA of 67 bases and, as in the case of msDNAs from myxobacteria (Yee, 1984; Dhundale, 1987), it can form a secondary hair-pin structure. The primary sequence, however, is not homologous to any of the myxobacterial msDNAs, nor to the msDNA from E. coli B (msDNA-Ec86; Lim and Maas, personal communication).

The sequence of the RNA molecule was determined using the RNA-DNA complex purified from E. coli Cl-1. The RNA sequence was determined using base specific RNases as described previously (Dhundale *et al.*, 1988). As shown in Figure 8, a large gap is observed in the RNA sequence "ladder". This gap is due to the DNA strand branched at the 2' position of the 15th rG residue of the RNA strand which produces a shift in mobility of the sequence ladder (see Figure 7). The RNA consists of 58 bases with the DNA molecule branched at the G residue at position 15 by a 2',5'-phosphodiester linkage. The branched G structure was determined as previously described for msDNAs from myxobacteria (Dhundale, 1987; Furuichi *et al.*, 1987). After RNase (A and T1) treatment, msDNA retains a small oligoribonucleotide linked to the 5' end of the DNA molecule due to the inability of RNases to cleave in the vicinity of the branched linkage. The 5' end was labeled with [γ -³²P]ATP using T4 polynucleotide kinase and the labeled RNA molecule was detached from the DNA strand by a debranching enzyme purified from HeLa cells (Ruskin *et al.* 1985; Arenas *et al.*, 1987; the debranching enzyme was a gift from Jerard Hurwitz). This small RNA was found to be a tetraribonucleotide which could be digested with RNase T1 to yield a labeled dinucleotide (not shown). Since RNase T1 could not cleave the RNA molecule at the G residue before debranching enzyme treatment, it was concluded that the single-stranded DNA is branched at the G residue via a 2',5'-phosphodiester linkage. In addition, partial RNase U₂ digestion cleaved the RNA molecule to yield a ³²P-labeled mono- and a ³²P-labeled trinucleotide (not shown). Thus, the sequence of the tetranucleotide is ^{5'}A-G-A-(U or C)^{3'}. Based on these data, the complete structure of msDNA-Ec67 from E. coli Cl-1 is presented in Figure 7. Despite a lack of primary structural homology, msDNA-Ec67 displays all the unique features found in msDNAs from myxobacteria. These include a single-stranded DNA with a stem-and-loop structure, a single-stranded RNA with a stem-and-loop structure, a 2',5'-phosphodiester linkage between the RNA and DNA, and a DNA-RNA hybrid at their 3' ends. This hybrid structure was confirmed by demonstrating sensitivity of the RNA molecule to RNase H (not shown).

Cloning of the locus for msDNA-Ec67

In order to identify the DNA fragment which is responsible for msDNA synthesis in E. coli Cl-1, Southern blot hybridization was carried out with various restriction enzyme digests of total chromosomal DNA prepared from E. coli Cl-1, using msDNA-Ec67 labeled with AMV-RT (the same preparation as shown in lane 3, Figure 6) as a probe. The result is shown in Figure 9A. EcoRI (lane 1), HindIII (lane 2), BamHI (lane 3), PstI (lane 4) and Bg11II (lane 5) digestions showed single band hybridization signals corresponding to 11.6, 2.0, ~22, 2.8 and 2.5 kilobase pairs (kb), respectively. The upper band appearing in the EcoRI digestion is due to incomplete digestion of the chromosomal DNA. Analysis of total chromosomal DNA prepared from E. coli Cl-1 by agarose gel electrophoresis revealed that the strain contains two plasmids of different size. However, neither plasmid hybridized with the ³²P- labeled probe, indicating that the fragments detected in Figure 9A are derived from chromosomal DNA. Furthermore, there is only one location for the msDNA-coding region on the chromosome, since various restriction enzyme digestions gave only one band of varying sizes. Similar results were observed for the msDNAs of myxobacteria (Yee et al., 1984; Furuichi et al., 1987; and Dhundale et al., 1988).

The 11.6-kb EcoRI fragment and the 2.8-kb PstI fragment were each cloned into pUC9 (Yanisch-Perron et al., 1985) and E. coli CL83 (a recA transductant of strain JM83), an msDNA-free K-12 strain (lane 1, Figure 9B), was transformed with the plasmids. Cells transformed with the 11.6-kb EcoRI clone (pCl-1E) were found to produce msDNA (lane 2, Figure 9B), whereas cells transformed with the 2.8-kb PstI clone (pCl-1P) failed to produce any detectable msDNA (lane 3, Figure 9B). A map of the 11.6-kb fragment is shown in Figure 10. Southern blot analysis of the fragment revealed that a 1.8-kb PstI - HindIII fragment hybridized with the msDNA probe. When the DNA sequence of this fragment was determined, a region identical to the sequence of the msDNA molecule was discovered. The DNA sequence corresponding to the sequence of msDNA is indicated by the enclosed box on the lower strand in Figure 11 and the orientation is from right to left. The location of this sequence is also indicated by a small arrow in Figure 10. As is the case for all other known myxobacterial msDNAs (Dhundale et al., 1987; Furuichi et al., 1987; and Dhundale et al.,

1988), a sequence identical to that of the RNA linked to msDNA (see Figure 7) was found downstream of the msDNA-coding region in opposite orientation and overlapping with that region by 7 bases. This sequence is indicated by the enclosed box on the upper strand in Figure 11 and the branched G residue is circled. Again, as in all the msDNAs found in myxobacteria, there is an inverted repeat comprised of a 13-base sequence immediately upstream of the branched G residue (residue 250 to 262; sequence a2 in Figure 11) and a sequence at the 3' end shown by an arrow in Figure 11 (residue 368 to 380; sequence a1). As a result of this inverted repeat, a putative longer primary RNA transcript beginning upstream of the RNA coding region and extending through the msDNA coding region would be able to self-anneal and form a stable secondary structure, which is proposed to serve as the primer as well as the template for msDNA synthesis (Dhundale *et al.*, 1987).

Existence of an essential gene for msDNA synthesis

The 2.8-kb PstI fragment (from PstI(a) to PstI(b) in Figure 10) was not able to synthesize msDNA. However, an overlapping 3.9-kb fragment from BamH (1.0 kb downstream of PstI(a); see Figure 10) to the following EcoRI site contains all the information required for synthesis of msDNA. This indicates that a region downstream of the PstI(b) site (Figure 10) is required for msDNA production. The nucleotide base sequence from this region revealed a long open reading frame (ORF) of 586 amino acid residues, starting with the initiation codon ATG at nucleotide 418 to 420 as shown in Figure 11. A distance of only 51 bases separates the initiation codon from the region which encodes msDNA. A putative Shine-Dalgarno sequence (GGA) can be found 10 bases upstream of the initiation codon. When the lacZ gene was fused in frame at the HindIII site (within the ORF) at amino acid residue-126, β -galactosidase activity was detected (not shown). Thus the region encompassing the ORF is indeed transcribed and the gene product encoded by the ORF is essential for msDNA synthesis. In a preliminary experiment, both msdRNA and the ORF appeared to be transcribed as the same transcription unit, since a deletion mutation removing the sequence from residue 1 to 181 blocked the expression of the lacZ gene fused at the HindIII site. A putative promoter can be found in the deleted sequence as boxed in Figure 11. These -35 and -10 regions

probably serve as the promoter for both msdRNA synthesis and the ORF.

Sequence similarity with retroviral reverse transcriptases

When the amino acid sequence of the ORF was compared with known proteins, a striking similarity was found with retroviral RTs. In Figure 12, the ORF is compared with RTs from HIV (human immunodeficiency virus; Ratner *et al.*, 1985; and Johnson *et al.*, 1986), and HTLV1 (human T-cell leukemia virus type I; Seiki *et al.*, 1983; and Patarca *et al.*, 1984). The first domain (Asn-32 to Val-291) matches well with the RT domains of HIV and HTLV1. In particular, the sequences around the polymerase consensus "Asp-Asp" sequence (Toh *et al.*, 1983; and Geng *et al.*, 1985; boxed in Figures 11 and 12) are well conserved. Out of 260 amino acid residues in this domain, 44 and 38 residues are identical with HIV and HTLV1, respectively. Between HIV-RT and HTLV1-RT, there are 78 identical amino acid residues in this domain.

The pol gene of retroviruses is known to produce a protein consisting of RT and RNase H activities; the former at the amino-terminal and the latter at the carboxyl-terminal region of the pol gene product (Ratner *et al.*, 1985; Johnson *et al.*, 1986; Varmus, 1985; and Tanese *et al.*, 1988). These domains have been shown to be separated by a poorly conserved "tether" domain of approximately 160 to 190 amino acid residues (Ratner *et al.*, 1985; Johnson *et al.*, 1986). On the basis of the HIV sequence; the similarities (only identical amino acid residues) between HIV and HTLV1 are 29.5 and 16.8% for the RT domain and the tether domain, respectively. The similarities between HIV and msDNA are 16.9 and 10.3% for the RT domain and the tether domain, respectively. The similarities between HTLV1 and msDNA are 14.6 and 15.5% for the RT domain and the tether domain, respectively. These results indicate that in addition to the RT region, there are reasonable similarities in the tether domain between retroviruses and msDNA. An alignment of the RNase H domains also revealed that there are similarities between retroviruses and msDNA (15.7 and 17.4% with HIV and HTLV, respectively; see Figure 12). The similarity between HIV and HTLV1 in this region is 18.0%.

Cell extracts were prepared and assayed for the presence of RT activity associated with the production of msDNA as predicted from the amino acid homologies. Only the E. coli strain (C2110, polA) (Tanese et al., 1985; Tanese et al., 1986; E. coli strain C2110 (polA1⁻) was a gift from M. Roth and S. Goff) harboring the plasmid, pCl-1EP5, containing the msDNA ORF displayed RT activity (Figure 13). The polA strain was used to eliminate high background activity in the RT assay due to DNA polymerase I. No RT activity was detected in extracts containing the vector plasmid alone, or when the template-primer (poly rC-dG) was absent from the reaction mix (Figure 13). It is interesting to note that the PstI(b) site is located at amino acid residue-430, which is between the tether domain and the RNase H domain. A plasmid lacking sequences downstream of the PstI(b) site did not produce msDNA. This suggests that the RNase H domain may be essential for msDNA synthesis, or alternatively that PstI disruption may result in inactivation of RT.

In addition to the similarity between msDNA-Ec67 RT and retroviral RT, there is an interesting similarity between msDNA and retroviruses; DNA synthesis starts at a site upstream of the RT-RNase H gene, and the orientation of DNA synthesis is opposite to the direction of transcription of the RT-RNase H gene. In the case of retroviruses, tRNAs are recruited from the cell for the priming reaction (Weiss et al., 1985), whereas for msDNA an RNA transcript serving as, template also serves as a primer by self-annealing to form a stable secondary structure (Dhundale et al., 1987; Furuichi et al., 1987).

Origin of the E. coli Reverse Transcriptase

At present the relationship between msDNA and retroviruses is an open question. It is possible that the study of msDNA may shed light on the question of the origin and evolution of retroviruses. It is an intriguing question to consider why some of the clinical E. coli strains, isolated from human patients produce msDNA. Our preliminary data indicate that msDNAs produced by four independent E. coli strains, isolated from urinary tract infections, share little homology. This suggests that there may be enormously large numbers of species of msDNA in E. coli. In contrast to msDNAs found in E. coli, msDNA-Mx162 from M. xanthus is highly conserved, since nine

independent M. xanthus strains isolated from various sites have msDNA which hybridizes with the original msDNA-Mx162 (Dhundale et al., 1985). Furthermore, msDNA from another myxobacterium, S. aurantiaca (msDNA-Sa163; Furuichi et al., 1987), also shows a high degree of homology to msDNA-Mx162 (Furuichi et al., 1987).

5 Several lines of evidence suggest that the RT gene found in the E. coli strain Cl-1 is not likely to have originated in E. coli, but rather was recently acquired from some other source. For example, only about 4% of E. coli strains tested were found to produce msDNA. In addition, the RT gene from strain Cl-1 does not cross hybridize to chromosomal DNA from four other E. coli strains which produce msDNA molecules, indicating that there is extensive diversity among these RT genes.

10 In contrast, a DNA fragment from the E. coli-K-12 sigma factor gene can hybridize to chromosomal DNA from all five msDNA producing, E. coli strains, indicating the conserved nature of sigma factors. An analysis of the E. coli RT gene indicates that the codon usage for this gene is remarkably different from most E. coli proteins. In particular, AGA and AGG, the least frequently (2.7%) used codons for arginine among 199 E. coli genes (Maruyama et al., 1986), occurs at a frequency of 64.5% in the E. coli RT gene. Similarly, CUG is the most commonly used codon for leucine (61.3%; Maruyama et al., 1986) in E. coli genes, while its prevalence in the RT gene is only 9.1%. The AT base pair content of the E. coli RT gene was calculated to be 67.6%, which is substantially higher than the AT content of the E. coli genome (45%; Fasman, 1976). The AT contents of HIV and HTLV1 RT genes are 62.1% and 47.8%, respectively. These facts pose an intriguing question as to how and when 20 the RT gene, as well as the msDNA coding region, were integrated into the genome of the clinical strain.

There are many questions to be answered, including (a) are there any particles associated with msDNA, (b) is the msDNA region transposable like the Ty element of yeast (Boeke et al., 1985; Eichinger et al., 1988), (c) can the element responsible for the production of msDNA be transferred from cell to cell, (d) can a RT from one strain (E. coli or myxobacteria) complement the 25 production of msDNA of other strains, (e) does the promoter for the RNA transcript have any similarities to the retroviral LTR, (f) are there any specific integration sites for the msDNA element

on the E. coli chromosome, (g) why is the branched G residue conserved, (h) is there an enzyme responsible for priming DNA synthesis at the 2'-OH position of the rG residue, (i) why and how does msDNA synthesis stop at one distinct site on the RNA template, and (j) how different biochemically are the msDNA RTs from retroviral RTs?

5 The existence of reverse transcriptase in prokaryotes, previously speculated upon (Dhundale *et al.*, 1987), is now evident. This fact raises intriguing questions concerning possible roles of this enzyme in the prokaryotes other than a role in msDNA production. Recently we also found that M. xanthus, in which msDNA was originally discovered, has a long ORF in the same manner as found for msDNA-Ec67. This ORF has a high degree of similarity to the E. coli RT. Since eight 10 independent isolates of M. xanthus produce homologous msDNA, the M. xanthus RT is likely to have been acquired at a very early stage of its evolution in contrast to the E. coli RT. The determination of the structures of both M. xanthus and other E. coli RTs will shed light on the key question of the origin of RT and its role in prokaryotes.

15 An important embodiment of the invention relates to the discovery of msDNA-producing retrons in a number of diverse bacterial groups. Thus, retron elements appear to be widely prevalent, at least amongst the purple bacteria or proteobacteria including *Proteus*, *Klebsiella* and *Salmonella* of the gamma subdivision; *Rhizobium* and *Bradyrhizobium* from the alpha subdivision; and *Nannocystis* (a myxobacterium) from the delta subdivisions. These are representatives of the three of the four major subdivisions of the purple bacteria of proteobacteria. 20 As shown above the retron-encoded RT is responsible for the synthesis of msDNAs.

25 The retron elements were discovered by detecting the presence of msDNA by one of two classic methods: the so-called "RT extension method", described by Lampson, B.C., M. Inouye and S. Inouye, 1991. Survey of multicopy single-stranded DNAs and reverse transcriptase genes among natural isolates of *Myxococcus xanthus*. *J. Bacteriol.* 173:5363-5370 and in Lampson, B.C., M. Viswanathan, M. Inouye and S. Inouye, 1990. Reverse transcriptase from *Escherichia coli* exists as a complex with msDNA and is able to synthesize double-stranded DNA. *J. Biol. Chem.* 265:8490-

LAW OFFICES
WEISER & ASSOCIATES
SUITE 500
230 SO. FIFTEENTH ST.
PHILADELPHIA, PA 19102
(215) 875-6383
TELECOPIER (215) 875-8394
8496 or polyacrylamide gel electrophoresis of a chromosomal DNA extract followed by staining with

ethidium bromide as described by Yee, T., T. Furuichi, S. Inouye, 1984. Multicopy Single-Stranded DNA Isolated from a Gram-Negative Bacterium, *Myxococcus xanthus*. Cell, Vol. 38, 203-209. Both of these publications are incorporated herein by reference. Both methods provide a reliable, convenient and conventional protocol for screening of bacteria for the presence of retron-encoded RT and msDNAs.

In accordance with the RT extension method, the DNA portion of msDNA is specifically ^{32}P radio labeled. Radio labeled from a total RNA preparation extracted from each bacteria strain to be screened. Twenty or more isolates of *proteus mirabilia*, *Klebsiella pneumoniae*, *Salmonella* species, *rhizobial* species, and *enterococcal* species were screened by this method. Low-molecular-weight bands (Fig. 20) indicated the presence of small labeled DNAs after polyacrylamide gel electrophoresis and autoradiography of the labeling reaction mixes. In addition, half of each labeling reaction mix was also treated with RNase A, causing a shift to a faster-migrating band, indicating that the labeled DNA is also associated with RNA. This is hallmark of the msDNA molecule as discussed above. Four of the 23 *P. mirabilia* isolates screened produced msDNA, while only 1 of 21 *K. pneumoniae* isolates and 4 of 70 *Salmonella* isolates screened produced msDNA. msDNA was detected in any of the 30 or so *enterococcal* strains screened by this method. It was concluded that the bacterial genera which contain msDNA producing retrons elements are representatives of three of the four major subdivision of the purple bacteria or Proteobacteria, as described above.

In accordance with this embodiment of the invention, it is noteworthy that the discovery of msDNA extends for the first time the distribution of retrons-elements to a new phylogenetic division of the purple bacteria, namely, the alpha subdivision. A collection of 63 *rhizobial* isolates (shown in Table 1) were screened for the presence of msDNA by the RT extension method. Among the 63 isolates, msDNA were detected in 10 (16% - Fig. 20 and Fig. 21). However, all 10 positive isolates give strong, clearly labeled bands with a typical shaft of a fast-migrating band after treatment with RNase A, indicating the presence of RNA and DNA in the labeled molecule.

The 10 retrons-encoding *rhizobial* strains include both fast growing (*rhizobium*) and slow-growing

(*Bradyrhizobium*) rhizobia.

The RT extension method comprises treating a preparation of total RNA, extracted from a bacterial strain to be tested, with RT from a suitable source in the presence of the deoxynucleotides dATP, dTTP, dGTP and dCTP, one of which is radiolabeled, e.g., [α - 32 P] dCTP, 5 electrophoresing the treated RNA preparation on a polyacrylamide gel and determining initially the presence or absence of msDNA in the bacterium of interest by detecting a band of radiolabeled DNA corresponding to the single-stranded DNA of msDNA. Typical examples of suitable sources of RT are avian myeloblastosis virus (AMV) and Moloney murine leukemia virus (Mo-MLV). Conceivably, the test could be automated.

10 Total RNA samples, which contain msDNA if present in the bacterium, are extracted from the bacterial strain of interest and prepared for RT extension as follows. Total RNA, prepared from a 5-ml culture from the bacterial strain, is added to 50 μ l of a reaction mixture containing: 50 mM tris-HCl (pH 8.3); 6 mM MgCl₂; 40 mM KCl; 5 mM DTT; 1 μ m dATP, dTTP and dGTP; 0.04 μ M dCTP; 0.2 μ M [α 32 P] dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture is incubated at 37⁰C for 30 minutes, then extracted with 50 μ l of phenolchloroform (1:1) and precipitated with ethanol. The samples are subjected to electrophoresis on a 4% acrylamide -8 M urea gel with appropriate nucleotide size markers, e.g., the Klenow fragment of DNA polymerase I. If the labeled sample is digested with ribonuclease (RNase) A before it is placed on the gel, a single band corresponding to single-stranded DNA is detected, which is indicative of the presence of msDNA. 20 An aliquot from each labeling reaction mixture is treated with 5 μ g of RNase for 10 minutes at 37⁰C just prior to electrophoresis to detect in the gel a shift to a faster - migrating species, indicating that each labeled DNA is also associated with RNA, which is the hallmark of the msDNA molecule.

Low-molecular weight bands in the gel indicate the presence of small labeled DNAs after polyacrylamide gel electrophoresis and autoradiography of the labeling reaction mixtures.

25 Multiple bands observed in some of the lanes of the gel even after RNase treatment may be due to incomplete extension by RT during the labeling reaction, or, alternatively, multiple forms or species of msDNA may exist in a given bacterium.

The Yee method for screening bacteria for the presence of retrons which synthesize msDNAs involves purifying by a conventional phenol extraction procedure total chromosomal DNA from the desired bacteria to be screened, electrophoresis on a five percent preparation acrylamide gel and checking for a satellite band. The major satellite band is cut out to extract the material in the 5 band to quantitate the material in the satellite band. Total chromosomal DNA is subjected to acrylamide gel electrophoresis, the gel is stained with ethidium bromide and densitometric scanning is employed to quantitate the satellite DNA against the pBR322 standard. The method is described in better details in Yee cited above.

A collection of rhizobial isolates from the United States Department of Agriculture 10 (USDA) Beltsville Rhizobium Culture Collection are screened for the presence of msDNA by the RT extension method. This collection represents isolates at different times, from different legume hosts and from different geographic locations. msDNAs are detected in 10 isolates. All 10 positive isolates give strong, clearly labeled bands of DNA, with a typical shift to a fast-migrating band after treatment with RNase A, indicating the presence of RNA and DNA in the labeled molecule. The 10 retron-encoding rhizobial strains include both fast-growing (Rhizobium) and slow-growing (Bradyrhizobium) rhizobia as follows: Rhizobium sp. (Acacia) 3002 and 3838, Bradyrhizobium sp. (Aeschynomene) 3516, Bradyrhizobium sp. (Albizia) 3004, Bradyrhizobium sp. (Erythrina) 3242, Rhizobium loti 3468 and 3503, Rhizobium trifolii 2048 and 2065 and Bradyrhizobium sp. (Vigna) 3447. See Figure 21

20 Total DNA from each of eight msDNA-producing strains clearly cross-hybridizes with a nod YAB (1.6 - kb Eco RI fragment) gene probe derived from Bradyrhizobium japonicum, confirming that these strains are members of the Rhizobiaceae.

In view of the diversity of retron elements in prokaryotic populations, it is not excluded that msDNA synthesizing retrons would be found in bacteria living in alkaline 25 environments, such as in alkaline environments: Plectonema nostocorum, Flavobacterium spp. Agrobacterium spp. Bacillus spp. Ectothiorhodospira spp.; in acidic environments: Thiobacillus thermophilica and thiooxidans, Thermoplasma acidophilus, Sulfolobus acidocaldarius, Cuanidium

5 caldarius, Bacillus acidocaldarius; in very high temperature environment (thermophilic): Sulfolobus acaidocaldarius, Caldariella acidophila, Thermus aquaticus; in very low temperature (psychrotrophic): Vibrio marinus, Pseudomonas spp., Cytophaga spp., Flavobacterium spp.; in high salt environments (halophilic): Halobacterium cutirubrum and salinarium, Halococcus morrhuae, Danaliella viridis; in high barometric pressure (like deep sea - barophilic), which are believed to inhibit the gut of ocean bottom dwelling fish. By using one of the two screening tests identified above, one skilled in the art will readily determine whether any one of these bacteria contain retrons synthesizing msDNA. This may be particularly interesting for making evolutionary comparisons between homologous RT genes present in distantly related phytogenic strains.

10 A representative number of amino acid sequences of representative RTs were analyzed to determine similarities and differences. The following observations were made. The amino acid sequences of these bacterial RTs are shown in Figure 14. The individual nucleotide and amino acid sequences for each of the RTs are shown in Figures 2, 11 and 15 through 19.

15 From a comparison of these sequences, it is noted that there are 61 conserved positions in the RT domains as indicated by solid dots at the bottom of the sequences in Figure 14. It is further noted that all bacterial RTs possess the YXDD sequence. Several other residues are conserved including the LPQS sequence that is especially common in retroviral reverse transcriptases. The RT domains are divided into seven subdomains. For each subdomain, the consensus sequences for the seven bacterial RTs can be established, as shown at the bottom of the sequences in Figure 14. There 20 are 18 extra residues (except 26 residues for RT-Ec67) between subdomains 2 and 3, in which there is a reasonably good consensus sequence.

It has been noted that the RTs of the present invention possess a number of common conserved sequences of nucleotides and amino acid residues.

The most common conserved sequence of amino acid residues noted is as follows:

25 tyrosine, alanine or cysteine and two aspartic acid residues. This conserved sequence, common to all RTs of the present invention, is also known as the YXDD sequence. *as shown in Seq. ID No. 4350*

A second conserved sequence of amino acid residues noted is as follows: serine, x which is a hydrophobic residue selected from the group consisting of valine, phenylalanine leucine and isoleucine, x_1 which is a polar residue selected from the group consisting of threonine, asparagine, lysine and serine and x_2 which is a hydrophobic residue selected from the group consisting of tryptophan, phenylalanine and alanine.

as shown in Seq. ID No. 4451

A third conserved sequence of amino acid residues noted is as follows: asparagine, x which is a hydrophobic residue selected from the group consisting of alanine, leucine and phenylalanine and x_1 which is a hydrophobic residue selected from the group consisting of leucine, valine and isoleucine.

as shown in Seq. ID No. 4452

10 A fourth conserved sequence of amino acid residues further noted is as follows: x which is a polar residue selected from the group consisting of arginine, glutamic acid, lysine, valine and glutamine, a second residue which is valine, a third residue which is threonine and a fourth residue which is glycine.

as shown in Seq. ID No. 4452

15 These conserved sequences are only a portion of the total number of common sequences of the RTs. For other conserved sequences held in common by the bacterial RTs reference is made to Figure 14.

20 The RTs of the other groups of bacteria described herein as capable of synthesizing msDNAs are likewise believed to have a similar profile of conserved nucleic acid and amino acid residue sequence similarities as shown in Figure 14 and discussed above. This observation also applies to the genus Nannocystis.

In accordance with the invention, it is contemplated that prokaryotic reverse transcriptase, which is essential for msDNA synthesis, may be responsible for host cell parasitic or selfish DNA synthesis. Additionally, it is thought that the prokaryotic reverse transcriptase molecule may be essential for synthesis of biological messengers and nucleic acid enzymes.

25 The msDNAs synthesized by the reverse transcriptase disclosed herein possess a highly stable RNA; it is capable of self-annealing and may serve as the primer and template for msDNA synthesis. The reverse transcriptases (RTs) disclosed herein may be used as diagnostic agents. It is

also contemplated that the RTs of the invention can synthesize msDNAs which will contain specific selected DNA fragments that can hybridize with complementary ssDNA, or otherwise identify ssDNAs, sought for, thus being useful as probes.

5 The possibility for the msDNAs to behave like restriction enzymes (or have restriction-like enzyme activity) in being capable of cleaving DNAs, or cut off a segment of itself, cannot be excluded.

The following examples are provided for purposes of illustration only and are not to be viewed as a limitation of the scope of the invention. The following examples are illustrative of bacterial isolates screened and identified to contain msDNA by way of the present invention.

10 **EXAMPLE 1**

One of the rhizobial strains, Rhizobium trifolii USDA 2065 is identified as containing msDNA by the RT extension method by which msDNA from total RNA is specifically labeled with ³²P as follows.

20 Total RNA from a 5-ml culture of R. trifolii 2065 is added to a 50 μ l reaction mixture containing: 50 mM tris-HCl (pH 8.3); 6 mM Mg Cl₂; 40 mM KCl; 5 mM DTT; 1 μ m dATP, dTTP and dGTP; 0.04 μ M dCTP; 0.2 μ M [α -³²P] dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture is incubated at 37⁰C for 30 minutes, then extracted with 50 μ l of phenolchloroform (1:1) and precipitated with ethanol. The samples are subjected to electrophoresis on a 4% acrylamide-8 M urea gel with appropriate nucleotide size markers, such as the Msp I digest of pBR322 end-labeled with [α -³²P] dCTP and the Klenow fragment of DNA polymerase I. An aliquot of the reaction mixture containing R. trifolii RNA is treated with 5 μ g of RNase for 10 minutes at 37⁰C prior to electrophoresis to detect in the gel a shift to a faster-migrating species, which indicates that the ³²P-labeled DNA extended by RT is also associated with RNA, which clearly demonstrates the presence of msDNA.

Low-molecular weight bands in the gel indicate the presence of small ^{32}P -labeled DNA after polyacrylamide gel electrophoresis and autoradiography. The labeled DNA is indicative of the presence of msDNA.

EXAMPLE 2

5 By the method described above in Example 1, (a) Proteus mirabilis 1174b is found to synthesize msDNA by the retrons containing the RT; (b) Klebsiella pneumoniae 912b is found to synthesize msDNA by RT; (c) Salmonella sp. strain SARB-3 is found to synthesize msDNA by the retrons containing the by the retrons containing the RT; (d) Nannocystis exedens Nael is found to synthesize msDNA by RT; (e) Bradyrhizobium spp. 3447, 3516 and 3004 are also found to synthesize msDNA by the retrons containing the RT.

The following method, exemplified for E. coli, for the isolation and purification of bacterial RT is applicable to bacteria which are screened as positive for the presence of msDNA by the RT extension in vitro method.

EXAMPLE 3

Isolation and Purification of Bacterial Reverse Transcriptase.

The following is a description of a convenient method for isolating and purifying a bacterial RT.

From 10 liters of a stationary phase culture of E. coli strain C2110 harboring plasmid pCl-1EP5b, cells are harvested, washed in 50 mM Tris (pH 8.0), and resuspended in lysozyme buffer

20 (50 mM Tris (pH 7.5), 10% sucrose, 0.3 M NaCl, 1 mM EDTA, 1 mM phenylmethylsulfonyl fluoride).

Fresh lysozyme is added to a final concentration of 2 mg/ml. The suspension is incubated on ice for 15 minutes followed by a quick freeze at -70°C , then thawed on ice. Lysis is enhanced by the addition of 2 volumes of buffer M (50 mM Tris (pH 7.0), 1 mM dithiothreitol, 0.2% Nonidet P-40,

10% glycerol, and 25 mM NaCl) followed by incubation on ice, then a quick freeze-thaw. A cleared lysate is obtained by centrifugation at 38,000 rpm in a 50Ti rotor for 30 minutes. The cleared lysate is fractionated by ammonium sulfate precipitation (0-50%, 50-70% and 70-90%), followed by dialysis overnight (4⁰C) for each fraction against buffer M. Ammonium sulfate fractions, 50-70% and 70-90%, show RT activity and are pooled, then applied to a DEAE-column (2.5 x 50 cm; DE52 Whatman) equilibrated with buffer M. The DE52 column is washed, and RT activity is eluted from the column at a range of 300 to 350 mM NaCl. The DE52 fractions showing RT activity are pooled, concentrated by membrane ultrafiltration (Amicon) and then loaded onto a Sephadryl S-300 column (Pharmacia LKB Biotechnology Inc., 1.5 x 75 cm) equilibrated with buffer M. The column is developed with the same buffer. Again, fractions from the S-300 column having RT activity are pooled and concentrated, and 0.7 ml is loaded onto a 16-30% glycerol density gradient. The glycerol gradients are set up and run as described previously (Viswanathan *et al.*, 1989). The purified Ec67.RT (fractions 7, 8 and 9) is stored as separate glycerol fractions at -20⁰C.

When this protocol is applied to the msDNA bacterial synthesizing strains, the respective RTs are isolated and identified as shown above.

Another convenient method for isolating and purifying reverse transcriptase is published in Lampson B.C., S. Inouye and M. Inouye, "msDNA of Bacteria", Progress in Nucleic Acid Research and Molecular Biology, Vol. 40, pages 1 *et seq.*

The invention has been described in detail with particular reference to the above embodiments. It will be understood, however, that variations and modifications can be affected within the spirit and scope of the invention.

REFERENCES

Birnboim H.C., and J. Dolly, Nucl. Acid Res., 7, 1513-1523 (1979).

Boeke J.D., Gorfinkel C.A., Styles C.A., Fink G.R., Cell, 40, 491 (1985).

Cairns J., Overbaugh J., Miller S., Nature, 335, 142-145 (1988).

Churchward G., Belin D., Nagaime Y., Gene, 31, 165 (1984).

Clare J., Farabaugh P., Proc. Natl. Acad. Sci. USA, 82, 2829-2833 (1985).

Dhundale A., Furuichi S., Inouye S., Inouye M., J. Bactériol., 164, 914 (1985).

Dhundale A., Lampson B., Furuichi T., Inouye M., Inouye S., Cell, 51, 1105 (1987).

Dhundale A., Inouye M., Inouye S., J. Biol. Chem., 263, 9055 (1988).

Dhundale A., Furuichi T., Inouye M., Inouye S., J. Bacteriol., 170, 5620-5624 (1988a).

Dickson C., Eisenman R., Fan H., Hunter E., Teich N., Molecular Biology of Tumor Viruses, ed. 2, Cold Spring Harbor Laboratory NY, 513, 648 (1982).

Eichinger D.J., Boeke J.D., Cell, 54, 955-966 (1988).

Fasman G., CRC Handbook of Biochem. and Mol. Biol., Nucleic Acids, Vol 2., 102 (1976).

Furuichi T., Dhundale A., Inouye M., Inouye S., Cell, 48, 47-53 (1987a).

Furuichi T., Inouye S., Inouye M., Cell, 48, 55-62 (1987b).

Hsu M.Y., Inouye S., Inouye M., J. Biol. Chem., 264 (1989).

Inouye S., Franceschini T., Inouye M., Proc. Natl. Acad. Sci., USA, 80, 6829-6833 (1983).

Inouye, S., Hsu, M.Y., Eagle, S. and Inouye, M., Cell, 56: 709-717 (1989).

Inouye, S., Herzer, P.J. and Inouye, M., Proc. Natl. Acad. Sci, 87: 942-945 (1990).

Johnson M.S., McClure M.A., Feng D.F., Gray J., Doolittle R.F., Proc. Natl. Acad. Sci., USA, 83, 7648-7652 (1986).

Kaiser D., Ann. Rev. Genet, 20, 539-566 (1986).

Komano T., Franceschini T., Inouye S., J. Mol. Biol., 196, 517-524 (1987).

Lampson B.C., Inouye M., Inouye S., Cell, 56, 701-707 (1989a).

Lampson B.C., Inouye S., Inouye M., "msDNA of Bacteria", Progress in Nucleic Acid Research and Molecular Biology, Vol. 40, pages 1 et seq.

Lampson B.C., Sun J., Hsu M.Y., Vallejo-Ramirez J., Inouye S., Inouye M., Science, 243, 1033-1038 (1989b).

Lampson B.C., Viswanathan M., Inouye M. and Inouye S., "Reverse Transcriptase from *Escherichia coli* Exists as a Complex with msDNA and is Able to Synthesize Double-stranded DNA", J. Biol. Chem. 265: 8490- 8496 (1990)

Lim, D. and Maas, W.K., Cell, **56**: 891-904 (1989).

Lim D., Maas W., *Mol. Microbiol.*, 3 1141-1144 (1989).

Maniatis T., Fritsch E.F., Sambrook J., Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory, NY (1982).

Maruyama T., Gojobori T., Aota S., Ikemura T., Nuc. Acid Res., 14, r151-r189 (1986).

Maxam A.M., Gilbert W., *Meth. Enzymol.*, **65**, 499 (1980).

Nargang F.E., Bell, J.B., Stohl L.L., Lambowitz A.M., Cell, **38**, 441-453 (1984).

Perron Y.C., Vieria J., Messing J., Gene, 33, 103 (1985).

Ratner L., et al., *Nature*, 313, 277-283.

Rice N.R., Stephens R.M., Burny A., Gilden, R.V., *Virology*, 142, 357-377 (1985).

Romeo J.M., Emson B., Zusman D.R., Proc. Natl. Acad. Sci. USA, 83, 6332-6336 (1986).

Beth M J, Tanese N, Goff S P, J Biol Chem., 260, 9326 (1985).

Ruskin B., Green M., *Science*, 229, 4274 (1987).

Saigo K., Kugimiya W., Matsuo Y., Inouye S., Yoshioka K., Yuki S., Nature, **312**, 659-661 (1984).

Sanger F, Nicklen S, Coulson A R. Proc. Natl. Acad. Sci., USA, 74, 5463-5467.

Seiki M., Hattori S., Hirayama Y., Yoshida M., Proc. Natl. Acad. Sci. USA, **80**, 3618-3622 (1983).

Shippick T M, Lerner R A, Sutcliffe J G, Nature 293 543-548 (1981).

Steynshagen, J.B., Robins, D.M., *Cell*, 55, 247-254 (1988).

Southern E. J. Mol. Biol. 98: 503 (1975)

Tanasa N, Roth M, Goff SP. Proc Natl Acad Sci USA 82, 4944 (1985).

Tanasa N, Sedrocki J, Haseltine W A, Goff S P. J Virol 59: 743 (1986).

Tamm H.M. Cell 31: 599-600 (1980)

T. I. H. Hayashida H. Miyata T. Nature 305 827 (1983)

Vaccinia HE Nature 314 584-585 (1985)

Vidrio, J., Messing, J., *Gene*, 19, 259-268 (1982)

Visawanathan M., Inouye M., Inouye S., J. Biol. Chem., 264, 13665-13671 (1989).

Voytas D.F., Ausbel F.M., Nature, 336, 242-244 (1988).

Weiss N., Teich H., Varmus H., Coffin J., RNA Tumor Viruses, Vol. 2, Cold Spring Harbor Laboratory (1985).

Yee T., Furuichi T., Inouye S., Inouye M., Cell, 38, 203-209 (1984).

Yuki S., Ishimaru S., Inouye S., Saigo K., Nucl. Acid Res., 14, 3017-3020 (1986).

269050 - 1000000000